**37252 Regression and Linear Models**

**Lab 5: Multiple Linear Regression III**

This lab is marked out of 21.

Please save your file in PDF format with name

**37252\_Lab5\_Surname\_FirstName**

**Due: 12 noon Wednesday 10 April 2024**

In this week’s lab we model the calorie content of breakfast cereals sold in a particular supermarket. The data are nutritional information from 77 products and available in **37252\_Lab5\_data.csv**, which can be downloaded from Canvas.

The variables we consider are summarised in the table below.

|  |  |  |
| --- | --- | --- |
| **Name** | **Role** | **Description** |
| $$calories$$ | response | calorie count |
| $$sugars$$ | predictor | sugar content |
| $$carbo$$ | predictor | carbohydrate content |
| $$fat$$ | predictor | fat content |
| $$shelf$$ | predictor | display shelf (1, 2, 3) |

The variable $shelf$ is a three-state categorical variable which must be recoded into two dummy variables; we can do this either manually or ask R to do it for us by specifying it as a *factor*.

Build a multiple linear regression model with $calories$ as response and $sugars$, $carbo$, $fat$, $shelf1$ (dummy variable for $shelf=1$ ), $shelf2$ (dummy variable for $shelf= 2$), interaction between $shelf$ and $sugars$ as predictors.

>caloriesdat<-read.csv("~/2024\_37252/Labs/Lab5/37252\_Lab5\_data.csv")

> caloriesdat$shelf <- as.factor(caloriesdat$shelf)

> caloriesdat$shelf <- relevel(caloriesdat$shelf, ref = "3")

> mod1 <- lm(calories ~ sugars + carbo + fat + shelf + sugars\*shelf, data = caloriesdat)

> summary(mod1)

1. Write down the estimated regression equation **[1 mark]** and provide interpretations of the estimated beta coefficients for $shelf1$ and $sug\\_shelf1$ **[2 marks]**.

Call:

lm(formula = calories ~ sugars + carbo + fat + shelf + sugars \*

 shelf, data = caloriesdat)

Residuals:

 Min 1Q Median 3Q Max

-13.4877 -3.2205 -0.8144 2.5832 20.9621

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 16.3052 3.9262 4.153 9.24e-05 \*\*\*

sugars 4.3232 0.2819 15.334 < 2e-16 \*\*\*

carbo 3.6294 0.2211 16.414 < 2e-16 \*\*\*

fat 8.5000 0.7836 10.847 < 2e-16 \*\*\*

shelf1 5.5158 3.0066 1.835 0.0709 .

shelf2 5.8135 4.0466 1.437 0.1553

sugars:shelf1 -1.2745 0.4384 -2.907 0.0049 \*\*

sugars:shelf2 -1.2589 0.4458 -2.824 0.0062 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.934 on 69 degrees of freedom

Multiple R-squared: 0.9158, Adjusted R-squared: 0.9072

F-statistic: 107.2 on 7 and 69 DF, p-value: < 2.2e-16

**General**

$$\hat{calories}=16.305+4.323×sugars+3.629×carbo+8.5×fat$$

 $+5.516×shelf1+5.814×shelf2$

$$ -1.275×sugars×shelf1-1.259×sugars×shelf2$$

**[1 mark]**

**Shelf 1**

$$\hat{calories}=21.821+3.048×sugars+3.629×carbo+8.5×fat$$

**Shelf 2**

$$\hat{calories}=22.119+3.064×sugars+3.629×carbo+8.5×fat$$

**Shelf 3**

$$\hat{calories}=16.305+4.323×sugars+3.629×carbo+8.5×fat$$

The coefficient $\hat{β}\_{shelf1}=5.516$ is the predicted difference in $calories$ for shelf 1 cereals compared to shelf 3 cereals with the same $carbo$ and $fat$ and when $sugars=0$ **[1 mark]**.

The coefficient $\hat{β}\_{sug\\_shelf1}=-1.275$ is the predicted difference in the change in calories for a unit increase in $sugars$ holding $carbo$ and $fat$ steady for shelf 1 cereals compared to shelf 3 cereals **[1 mark]**.

**OR**

The coefficient $\hat{β}\_{sug\\_shelf1}=-1.275$ is the predicted difference in the in the sensitivity of $calories$ to $sugars$ for shelf 1 cereals compared to shelf 3 cereals **[1 mark]**.

1. Use R to calculate $\hat{calories}$ for cereals on shelves 1, 2 and 3 when $sugars=9$, $carbo=16$ and $fat=3$ **[3 marks]**.

> newdata <- data.frame(shelf = c("1","2","3"), sugars = 9, carbo = 16, fat = 3)

> predict(mod1, newdata)

 1 2 3

132.8299 133.2683 138.7848

**Shelf 1**

$$\hat{calories}\left(sugars=9,carbo=16,fat=3,shelf1=1,shelf2=0\right)$$

$$=132.83$$

**[1 mark]**

**Shelf 2**

$$\hat{calories}\left(sugars=9,carbo=16,fat=3,shelf1=0,shelf2=1\right)$$

$$=133.27$$

**[1 mark]**

**Shelf 3**

$$\hat{calories}\left(sugars=9,carbo=16,fat=3,shelf1=0,shelf2=0\right)$$

$$=138.78$$

**[1 mark]**

1. Determine if there is any statistical evidence of serial correlation which would violate the assumption independence **[2 marks]**.

> durbinWatsonTest(mod1)

 lag Autocorrelation D-W Statistic p-value

 1 0.3076085 1.344031 0.002

 Alternative hypothesis: rho != 0

The DW statistic of 1.344 is **[1 mark]**, and p value is < 0.05 hence there is evidence of serial correlation **[1 mark]**.

1. Determine if there is any statistical evidence of multicollinearity **[3 marks]**.

> library("rms")

> vif(mod1)

 sugars carbo fat shelf1 shelf2 sugars:shelf1

 3.355575 1.603708 1.342469 3.800307 7.101653 4.009479

sugars:shelf2

 9.899877

Note: we need to use vif function in this package for categorical variables.

With VIFs of 7.102 and 9.9 respectively (above the warning level of 5 and near the alarm level of 10) **[1 mark]**, there is evidence that $shelf2$ and $sug\\_shelf2$ are collinear **[1 mark]**.

Multicollinearity is to be expected with interaction terms **[1 mark]**.

1. Perform a visual analysis of the residuals for compliance with the normality, independence and constant variance assumptions **[3 marks]**.

 

 

 

Normality assumption – PP plot in particular shows large departures from normality **[1 mark]**.

Independence assumption – maybe a general increase in residuals against $fat$, so potential problem with this assumption **[1 mark]**.

Constant variance assumption – maybe a hint of changing variance in plot against $carbo$, so potential problem with this assumption **[1 mark]**.

1. Identify potentially influential points giving statistical evidence for your answer **[3 marks total, 1 for evidence, 2 for identification of all points]**.

> library('olsrr')

> ols\_plot\_cooksd\_bar(mod1)



The critical Cook’s D for this model is given by

$$D\_{critical}=\frac{4}{n}=\frac{4}{77}≈0.052.$$

**[1 mark]**

The data points with Cook’s D in excess of the critical value are records 3, 30, 33, 34 and 60 **[2 points]**.

Filter out the 5 points identified in part (f) (hint!) and re-run the regression model again (see Lab 2 for filtering instructions).

From part (e) you probably identified a problem with the assumption of normality (hint!). We’ll now test the normality of the residuals in the filtered dataset model.

1. Using 0.05 significance level, perform a hypothesis test as to the normality of the residuals (see Lab 2 for instructions). Write down the hypotheses **[1 mark]**, the test statistic and p-value **[1 mark]**, the result of the test **[1 mark]** and a conclusion in non-mathematical language **[1 mark]**.

> caloriesdat\_red <- caloriesdat[-c(3, 30, 33, 34, 60),]

> mod1\_red <- lm(calories ~ sugars + carbo + fat + shelf + sugars\*shelf, data = caloriesdat\_red)

> mod1\_red.st.resid<-rstandard(mod1\_red)

> shapiro.test(mod1\_red.st.resid)

 Shapiro-Wilk normality test

data: mod1\_red$residuals

W = 0.98672, p-value = 0.6512

**Hypotheses**

$H\_{0}:$ the residuals $\hat{ϵ}\_{i}$ are normally distributed

$H\_{A}:$ the residuals $\hat{ϵ}\_{i}$ are not normally distributed **[1 mark]**

**Test statistic and p-value**

The test statistic is $sw=0.989$ with p-value reported as $p=0.775$ **[1 mark]**.

**Test decision**

Retain null hypothesis as $p>0.05$ **[1 mark]**.

**Conclusion**

There evidence is not strong enough to conclude that residuals in the filtered dataset model are not normally-distributed **[1 mark]**.